DOCUMENT RESUME

ED 032 648 EA 002 555

By-Sirotnik, Kenneth

An Analysis of Variance Framework for Matrix Sampling.

California Univ., Los Angeles. Center for the Study of Evaluation.
Spons Agency-Office of Education (DHEW), Washington, D.C. Bureau of Research.
Report No-CSE-R-52
Bureau No-BR-6-1646-52
Pub Date May 69
Contract-OEC-4-6-061646-1909
Note-59p.
EDRS Price MF-$0.50 HC-$3.05
Descriptors-*Analysis of Variance, Bibliographies, Literature Reviews, Psychometrics, Research Methodology, *Sampling, *Statistical Analysis

Significant cost savings can be achieved with the use of matrix sampling in estimating population parameters from psychometric data. The statistical design is intuitively simple, using the framework of the two-way classification analysis of variance technique. For example, the mean and variance are derived from the performance of a certain grade level of students on arithmetic fundamentals. From the matrix of students and proposed test problems in the cell determined by the test and grade level variables, random sampling from both categories will provide efficient estimates of the mean and variance. Formulations for finite and infinite populations are derived. This document reduces the theoretical complexities to readable form: it includes a short literature review, a description of the technique using examples, an extension of the technique to multiple matrix sampling, and a discussion of the negative variance estimate problem using multiple matrix sampling. (LN)

CENTER FOR THE STUDY OF EVALUATION

Marvin C. Alkin, Director
Publications Committee:
James W. Trent, Chairman
Theodore R. Husek
Sherman J. Pearl
Audrey Schwartz

UCLA Graduate School of Education

The CENTER FOR THE STUDY OF EVALUATION is one of nine centers for educational research and development sponsored by the United States Department of Health, Education and Welfare, Office of Education. The research and development reported herein was performed pursuant to a contract with the U.S.O.E. under the provisions of the Cooperative Research Program.

Established at UCLA in June, 1966, CSE is devoted exclusively to finding new theories and methods of analyzing educational systems and programs and gauging their effects.

The Center serves its unique functions with an inter-disciplinary staff whose specialties combine for a broad, versatile approach to the complex problems of evaluation. Study projects are conducted in three major program areas: Evaluation of Instructional Programs, Evaluation of Educational Systems, and Evaluation Theory and Methodology.

This publication is one of many produced by the Center toward its goals. Information on CSE and its publications may be obtained by writing:

Office of Dissemination
Center for the Study of Evaluation
UCLA Graduate School of Education
Los Angeles, California 90024

AN ANALYSIS OF VARIANCE FRAMEWORK FOR MATRIX SAMPLING

Kenneth Sirotnik

CSE Report No. 52
May 1969

Center for the Study of Evaluation
UCLA Graduate School of Education
Los Angeles, California

ABSTRACT

This paper, a discussion of the methodology of matrix sampling, and the empirical and theoretical research on matrix sampling, attempts to demonstrate the following points:

1. Matrix sampling can be viewed as a simple two factor, random model analysis of variance design, the matrix sampling formulas for estimating the mean and variance being simply the point estimate formulas for estimating components of the underlying linear model.

2. These formulas can be based on the weakest possible set of assumptions, viz., random and independent sampling of examinees and items. No assumptions about the statistical nature of the data need be made.

3. The literature is unclear about what effect the above sampling assumptions have upon matrix sampling in the estimation of the mean and, especially, the variance.

4. Of the three alternative procedures suggested for dealing with negative variance estimates in multiple matrix sampling--equating the negative estimate to zero, Winsorizing the distribution of estimates, or treating all estimates alike regardless of sign--the third procedure appears to be most promising. A simulation study is necessary to determine the shape of the distribution of variance component estimates for matrix sampling as well as the relative efficiency of the three methods for handling negative estimates.

INTRODUCTION

Matrix sampling as a psychometric technique for estimating test score parameters is a relatively new technique. The most concise and complete discussion of this technique appears in Lord and Novick (1968). The theory used by Lord to derive the matrix sampling estimate formulas is, however, highly sophisticated and equally complicated. If matrix sampling were to become a sufficiently useful technique so as to warrant its inclusion into a less sophisticated, but more widely readable textbook on measurement theory (such as Gullikson, 1950; Horst, 1966; Magnusson, 1966), the Lord presentation would be undesirable from the standpoint of its complexity. The present formulation relies on the direct application of familiar point estimate procedures in the analysis of variance. Since such procedures are more widely known and have greater intuitive appeal, the author feels that they would be more amenable to the purposes of the "average" measurement text.

The material that follows is organized into four sections. The first two sections concentrate on describing the technique of matrix sampling (with examples) and reviewing most of the literature on the theoretical development and empirical validation of the technique. Section 3 presents the derivation of formulas for the mean and variance estimates using a relatively simple analysis of variance design. In section 4, the assumptions underlying the estimate formulas are discussed in relation to the use of multiple matrix

sampling. Emphasis is given to the negative variance estimate problem and procedures suggested to handle this problem.

1. THE MATRIX SAMPLING TECHNIQUE

Consider a large high school with, say, 250 students in the eleventh grade. Suppose the school administration decides for one reason or another that it is interested in knowing how proficient (defined in terms of the mean and variance) the eleventh grade is in, say, arithmetic fundamentals as measured by some test having, say, 30 arithmetic fundamental items.

Obviously, one approach would be to give all 250 students or examinees (denoted the <u>population of examinees</u>) the arithmetic test -- that is, each examinee would respond to all 30 problems or items (denoted the <u>population of items</u>). This would amount to 7500 (250 x 30) examinee-item responses. Depending upon how many examinees could take the test at one time and how long it would take to respond to each of the items, this testing could amount to a fairly long time -- more time, perhaps, than would be feasible given the schedules of the students, personnel, and school in general.

A second approach, one traditionally used in establishing norms for standardized tests, would be to <u>randomly</u> select a <u>sample</u> of examinees, say 125, and give them the entire 30-item test -- this procedure will be referred to as <u>examinee sampling</u>. Here, the sample of examinees' scores would be used to <u>estimate</u> what the mean and variance would have been had all 250 examinees taken the 30-item test.

A third approach, called <u>matrix sampling</u>, follows the procedures

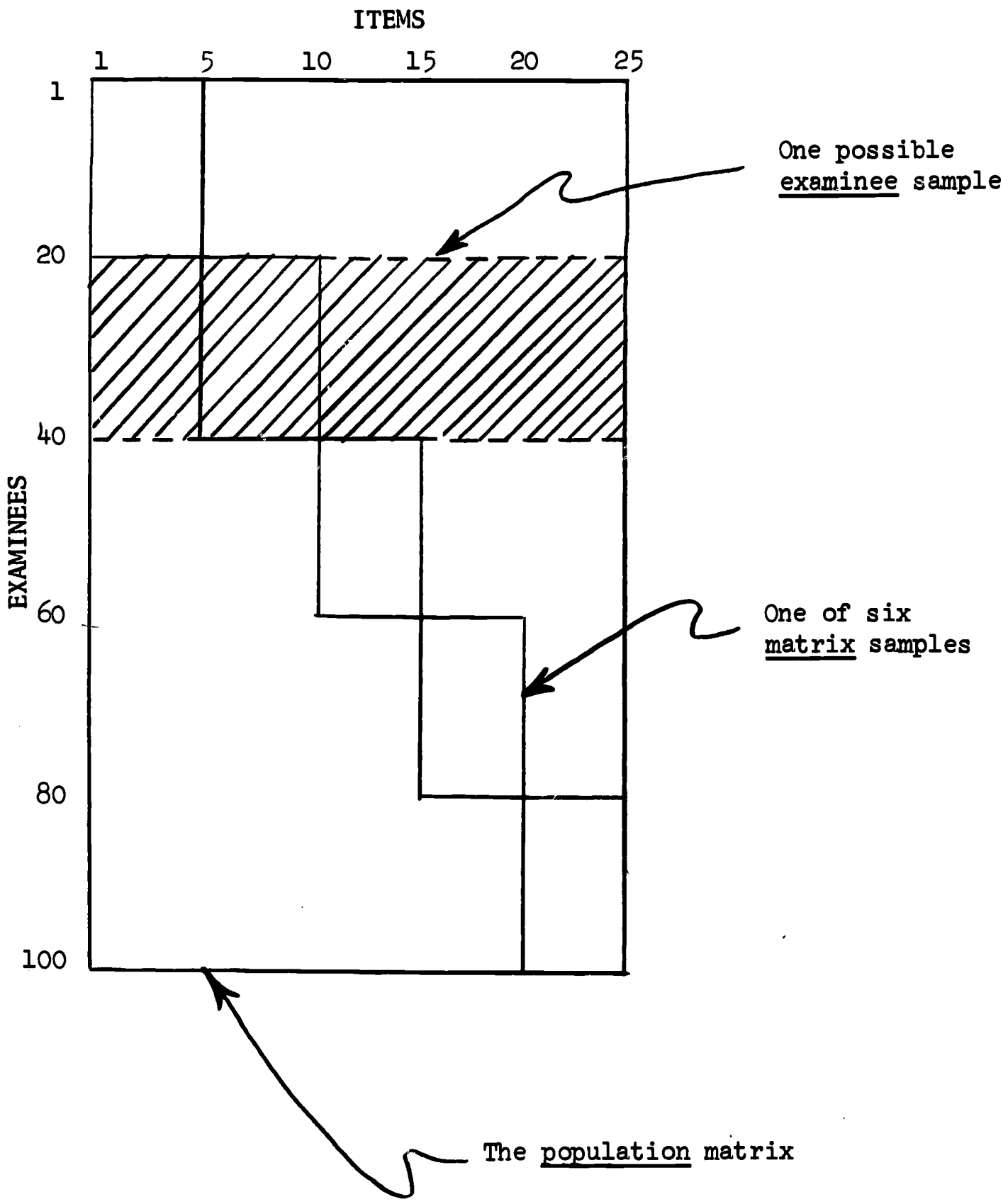
of the second approach, but with one important exception -- items, as well as examinees, are randomly sampled. For example, the sample of 125 examinees might each be given a sample of 15 items. Again, the data would be used to obtain estimates of what the mean and variance of the arithmetic fundamental scores would have been had the population of examinees responded to the population of items.

Now clearly, the first procedure, that of collecting complete data on everybody, would be most desirable. We would not have to estimate the mean score of all 250 examinees -- we could, in fact, compute the actual mean. The assumption that is being made here, however, is that the collection of complete data is not practical for various reasons, e.g., lack of time, money, personnel, etc.

If complete data are not obtained, then it would seem desirable that the sampling procedure employed sample as many items and examinees as possible. Thus, it would appear that examinee sampling is preferable to matrix sampling since the former sampled half of the examinee-item responses in the matrix population and the latter sampled only one quarter of these responses. However, if more than one matrix sample is strategically extracted from the population matrix -- a procedure called multiple matrix sampling -- matrix sampling can be more representative of the population than any other sampling procedure, given fairly stringent economical requirements. Figure 1 illustrates this point. The large rectangle represents the examinee-by-item population matrix of responses for a population of 100 (randomly arranged) examinees and 25 (randomly arranged) items.

FIGURE 1

Multiple Matrix Sampling *vs*. Examinee Sampling



The smaller rectangles, arranged diagonally, represent 5 random matrix samples having 20 examinees and 5 items each; thus, each matrix sample contains 100 examinee-item responses. The shaded rectangle represents one possible examinee sample of 20 examinees responding to all 25 items -- a total of 500 examinee-item responses. Clearly, both the combined matrix samples and the single examinee sample require the same number of examinee-item responses (one fifth of the matrix population). However, the matrix samples are more representative of the matrix population of examinee-item responses than is the examinee sample. Each of the 5 multiple matrix samples would yield estimates of the mean and variance of the arithmetic fundamental scores. The 5 mean estimates can be averaged to produce a final estimate of the mean; the 5 variance estimates can be averaged to produce a final estimate of the variance.

By way of summary, the fundamental methodological advance of matrix sampling is this: every examinee (from a finite or conceptually infinite population of examinees) need not respond to every item (from a finite or conceptually infinite population of items) in order to obtain estimates of the moments of the distribution of the population of examinees' responses to the population of items. This paper will be concerned only with the first and second moments, i.e., the mean and variance of the examinee score distribution. Analogous procedures can be used to estimate these parameters of the item "score" distribution.

From this description of the technique, it should be clear that the more popular name "item sampling" is a misnomer. It is not only

items that are being sampled, but examinees are being sampled as well. In other words, it is a two dimensional, examinee-by-item array of responses that is being sampled from the population examinee-by-item response array. For this reason, the older term "matrix sampling" (see review of Lord's initial papers in section 2) is used in this paper.

2. REVIEW OF THE LITERATURE

The research available regarding matrix sampling can generally be divided into two classes: that concerning the empirical validation of matrix sampling and that dealing with the theoretical development of matrix sampling. This chapter will briefly review these two classes of literature in that order.

Most of the empirical research on matrix sampling has consisted of studies attempting to verify that matrix sampling does what it is intended to do. These studies have followed one basic paradigm:

1. Obtain the entire matrix population of responses, thus obtaining the actual values of the population parameters to be estimated.

2. Generate parameter estimates using both the multiple matrix sampling and the more traditional examinee sampling methods.

3. Compare the matrix sampling estimates to those of the examinee samples in terms of closeness to the actual population values.

The first such study (Lord, 1962) employed a 70 item test and 1000 examinees. All 70,000 examinee-item responses were obtained, and the mean and variance of examinee test scores were calculated. Then 10 matrix samples of 7 items and 100 examinees each were randomly generated; the separate matrix sample estimates were averaged, yielding final estimates of the population mean and variance. Also, the 100

examinees in each sample were scored on all 70 items, creating 10 examinee samples; for each of these samples, mean and variance estimates were obtained in the usual manner. By comparison, the matrix sampling estimate of the mean was closer (in absolute difference) to the population mean than were 7 of the 10 examinee sampling estimates; the matrix sampling estimate of the variance was closer to the population variance than 5 of the 10 examinee sampling estimates. Lord points out that one reason why these results are not more strikingly in favor of matrix sampling is that the item samples were drawn <u>with replacement</u> -- that is, they were not nonoverlapping as a more efficient design would dictate.

Plumlee (1964) followed the basic paradigm with a 30 item test and 200 examinees. Although <u>nonoverlapping</u> matrix samples were used, the matrix sampling variance estimate was closer to the population value than only 1 of 10 examinee sampling estimates. Matrix sampling estimated the mean, however, better than all but 2 examinee samples.

Cook and Stufflebeam (1967 or 1967) extended the paradigm for validating the estimation procedures of matrix sampling by using variable sized matrix samples on both the item and examinee dimensions. Their results generally support the findings of the above studies. Again, the variance was not as well estimated as the mean by matrix sampling.

Husek and Sirotnik (1967) followed the above paradigm but with two different kinds of tests: an achievement test designed to maximize variability among subjects and an objective-meeting test designed to

minimize variability among subjects. For the achievement test data, matrix sampling was more efficient without exception than the examinee samples. For the objective-meeting test, matrix sampling estimated the mean better than 4 and the variance better than 3 out of 5 examinee samples. It was concluded that the efficiency of matrix sampling might be dependent upon the purpose for which the test was intended.

Cahen, et al. (1967) used a different design to investigate the efficiency of matrix sampling estimates. Matrix sampling data (not the matrix population) was collected initially for a 50-item test on the first day of testing. On the second day, the entire matrix population of data was obtained for a "nominally" parallel 50-item test. Comparisons were then made between the matrix sampling estimates of the mean (variance estimates were not considered) of the first day with the population mean obtained on the second day. Discrepencies were discussed in relation to varying testing time limits and examinee sample sizes.

The theoretical literature will now be considered. The concept of matrix sampling as a psychometric technique for estimating population score parameters from partial data apparently originates with Fredrick Lord. In a series of five publications, Lord discussed the technique under several different names and within different but related contexts.

In 1955, Lord referred to matrix sampling as Type 12 sampling, a logical extension of Type 1 sampling (the sampling of examinees) and Type 2 sampling (the sampling of test items), with primary emphasis

on resulting standard errors of measurement. Matrix sampling was the term used by Lord (1959a) wherein the main concern was with the estimation of various moments of the distribution of examinee true scores and the relationship of true scores to observed scores. The term item sampling was introduced by Lord (1959b) referring again to the same process of matrix sampling, which was discussed as one of several possible true score models available in mental test theory. In 1960, Lord showed how the "item sampling" model could be used to estimate (a) true score distributions on lengthened and shortened forms of a given test and (b) the relationship between observed scores on two parallel test forms given data on only one form. Finally, Lord (1965) discussed the concept of matrix sampling explicitly in terms of a data gathering procedure. Most recently (1968), this last work has been revised and incorporated as a comprehensive chapter on "item sampling" in Lord and Novick's text on mental test theory.

In order to present matrix sampling in a rigorous and generalized framework, Hooke (1956a, 1956b) developed an algebra involving symmetric polynomials of the elements in a matrix. The functions are called generalized symmetric means (gsm's) and have the property of being inherited on the average, i.e., the expected value of a gsm in a matrix sample is equal to the same quantity in the matrix population. Certain linear combinations of gsm's, called bipolykays, turn out to be estimates of the moments of the matrix population, thus providing a convenient way to obtain formulas for the estimated examinee score mean and variance from a matrix sample. Appendix 3 contains a brief

introduction to Hooke's formulation and how it is used by Lord (1965 or Lord and Novick, 1968) to derive the following formulas for the estimated mean and variance (see Appendix 1 for notational definitions):

(1) $\hat{\mu} = \bar{X}$ for finite and infinite matrix populations

(2) $\hat{\sigma}^2_\lambda = \frac{n(N-1)}{NM(n-1)(m-1)} [m(M-1)s^2_y - (M-m)\{\bar{y}(1-\bar{y}) - s^2_p\}]$ for finite matrix populations

(3) $\hat{\sigma}^2_\lambda = \frac{n}{(n-1)(m-1)} [ms^2_y - \bar{y}(1-\bar{y}) + s^2_p]$ for infinite matrix populations .

3. ANALYSIS OF VARIANCE FORMULATION OF MATRIX SAMPLING

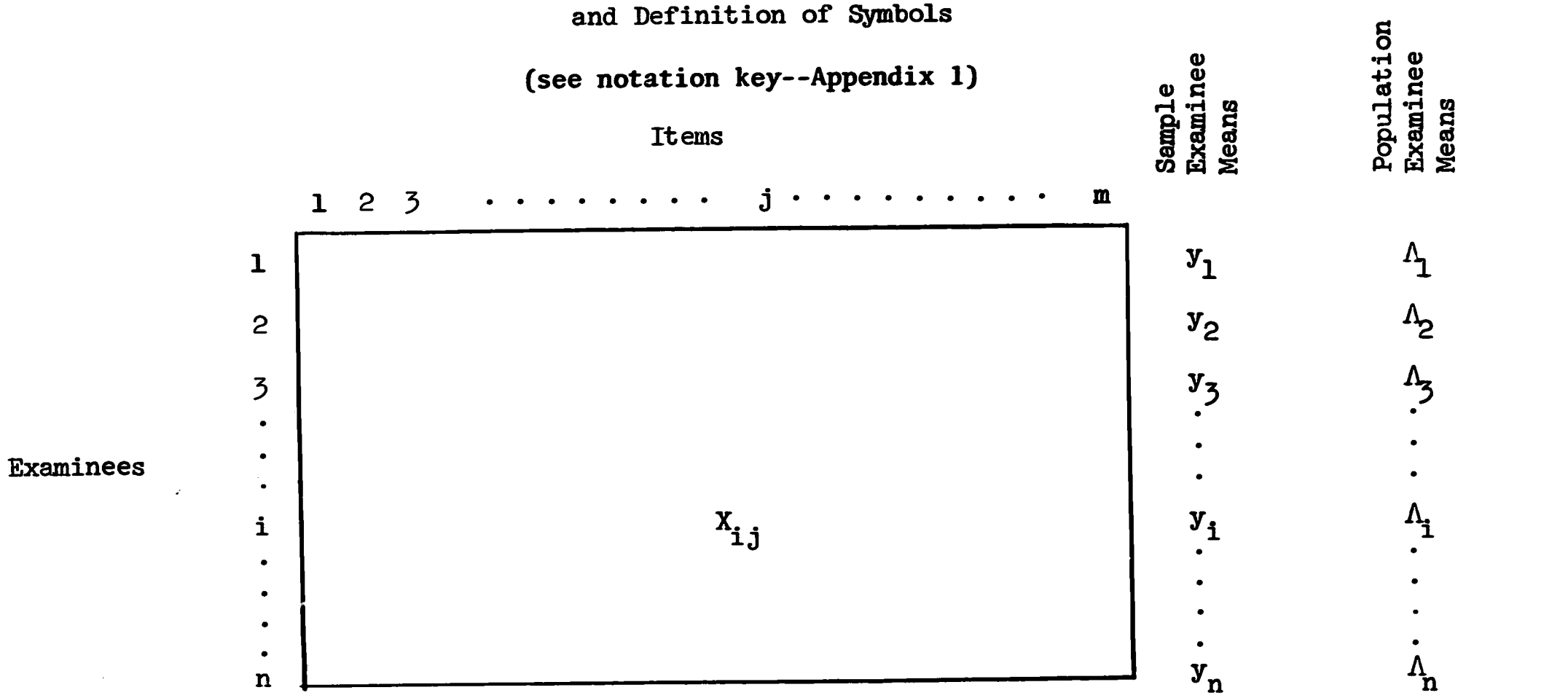
The main purpose of the present paper is to derive the formulas (1, 2, and 3) given at the end of the previous chapter using a simple examinees-by-items, repeated measures analysis of variance design. Although the use of analysis of variance procedures to obtain first and second moment estimates has been suggested by Lord and Novick (1968), the above formulas have never been explicitly derived without the use of bipolykays. Since Hooke's formulation is relatively complex and difficult to readily follow, the author feels that the subsequent presentation provides a convenient and simple exposition of the actual use of matrix sampling with perhaps a more intuitive feeling for the above formulas. Also, certain empirical problems resulting from the use of multiple matrix sampling, not explicitly clear in Lord's presentations, appear to be more easily discussed in terms of the present framework.

The two dimensional array in Figure 2 (taken in conjunction with the notation given in Appendix 1) defines the quantities which are used below. This array represents a typical matrix sample of n examinees and m items drawn independently and at random from corresponding populations of N examinees and M items. This array can also be considered as an $n \times m$ factorial design with one observation per cell. That is, an n-level examinee factor (E) is completely

FIGURE 2

Examinee x Item Matrix Sample Design
and Definition of Symbols

(see notation key--Appendix 1)



crossed (measured repeatedly) over the m-level item factor (I). The observation in any given cell represents an examinee-item response which can be either binary (0 or 1) or nonbinary (1, 2, 3, etc.). Furthermore, E and I are random factors, i.e., the levels (examinees) of factor E as well as those (items) of factor I are randomly selected from corresponding populations of levels. (N and M, the corresponding population sizes, can be either finite or infinite, depending upon the model viewed most reasonable by the researcher.)

In order to deal with the various sources of variability in this design, the following linear and additive model is traditionally used (Winer, 1962):

$$X_{ij} = \mu + \lambda_i + \pi_j + \varepsilon_{ij} \qquad (4)$$

where

μ = general level effect equivalent to the matrix population mean

$\lambda_i = (\Lambda_i - \mu)$ = examinee i effect in the population of examinees

$\pi_j = (\Pi_j - \mu)$ = item j effect in the population of items

$\varepsilon_{ij} = (X_{ij} - \lambda_i - \pi_j + \mu)$ = residual effect assumed to be due only to error of measurement.

In partitioning the total variability of the X_{ij} (s_x^2) into sources corresponding to the components in (4) and deriving their expected values (see Table 1), the following assumptions are made:

i) λ_i, π_j, and the ε_{ij} are independent, random variables with means of zero and variances of σ_λ^2, σ_π^2, σ_ε^2.* (Note homogeneity of error variances.)

ii) In view of the additivity of the model (no interaction effect), homogeneity of covariance is assumed among the population of error-free items. (This is equivalent to assuming that the item intercorrelation or covariance matrix has rank 1 -- with the exception of those populations of items where the intercorrelations consist <u>only</u> of <u>both</u> perfect positive (+1) and perfect negative (-1) correlations.)

Our main concern is to obtain estimates of μ and σ_λ^2, denoted $\hat{\mu}$ and $\hat{\sigma}_\lambda^2$, from the matrix sample data. (The estimate of item mean variance σ_π^2 can be obtained using analogous procedures.) This can be done as usual, selecting the appropriate E[MS] and solving for the desired variance component using the appropriate MS estimates.

For the estimate of μ, the expected value of both sides of (4) is taken as follows:

$$E[X_{ij}] = E[\mu + \lambda_i + \pi_j + \varepsilon_{ij}]$$

$$= \mu + 0 + 0 + 0 \quad \text{(by assumptions in i).}$$

*It should be emphasized that the assumption that these effects are normally distributed is not necessary in order to derive component of variance estimates.

TABLE 1

Analysis of Variance Source Table for Finite Factors
Including Degrees of Freedom (df), Sum of
Squares (SS), Mean Squares (MS),
and Expected Mean
Squares (E[MS])

Source	df		MS	E[MS]
E	n-1	$SS_E = m \sum (y_i - \bar{X})^2$	$MS_E = SS_E/(n-1)$	$(1 - \frac{m}{M})\,\sigma_\varepsilon^2 + m\sigma_\lambda^2$
I	m-1	$SS_I = n \sum (p_j - \bar{X})^2$	$MS_I = SS_I/(m-1)$	$(1 - \frac{n}{N})\,\sigma_\varepsilon^2 + n\sigma_\pi^2$
Residual	(n-1)(m-1)	$SS_R = \sum\sum (X_{ij} - y_i - p_j + \bar{X})^2$	$MS_R = SS_R/(n-1)m-1)$	σ_ε^2

Hence,

$$(5) \qquad \hat{\mu} = \bar{X} ,$$

that is, the mean of the matrix sample is an unbiased estimate of the mean in the matrix population, which is the same as formula (1) above, given by Lord.

To estimate σ_λ^2, the SS in Table 1 are first converted into the more familiar matrix sample statistics as follows:

(6) $SS_E = nms_y^2$ where s_y^2 is the sample variance of examinee mean scores

(7) $SS_I = nms_p^2$ where s_p^2 is the sample variance of item mean scores

(8) $SS_R = nm\bar{s}_j^2 - nms_y^2$ where $\bar{s}_j^2$ is the average item variance in the sample.

(Proofs of these equivalencies are found in Appendix 2, Proofs 1, 2, and 3.) We can now solve for the desired component of variance estimate $\hat{\sigma}_\lambda^2$. From Table 1 it is evident that

$$(9) \qquad \sigma_\lambda^2 = \frac{E[MS_E] - (1 - \frac{m}{M}) E[MS_R]}{m} .$$

Using corresponding MS as estimates, we have

$$(10) \qquad \hat{\sigma}_\lambda^2 = \frac{SS_E}{m(n-1)} - \frac{(1 - \frac{m}{M}) SS_R}{m(n-1)(m-1)}$$

Making substitutions using the above equivalencies (6 and 8) for the SS, we have

$$(11)\quad \hat{\sigma}^2_\lambda = \frac{nms_y^2}{m(n-1)} - \frac{(1 - \frac{m}{M})(nm\bar{s}_j^2 - nms_y^2)}{m(n-1)(m-1)}$$

$$= \frac{nm(m-1)s_y^2 - nm\bar{s}_j^2 + nms_y^2 + \frac{nm^2}{M}\bar{s}_j^2 - \frac{nm^2}{M}s_y^2}{m(n-1)(m-1)}$$

$$= \frac{nM(m-1)s_y^2 - nM\bar{s}_j^2 + nMs_y^2 + nm\bar{s}_j^2 - nms_y^2}{M(n-1)(m-1)}$$

$$= \frac{ns_y^2[M(m-1) + M - m] - n\bar{s}_j^2(M-m)}{M(n-1)(m-1)}$$

$$= \frac{n}{M(n-1)(m-1)}\,[m(M-1)s_y^2 - (M-m)\bar{s}_j^2]$$

At this point it must be noted that the use of sigma in σ^2_λ (and σ^2_π) is valid only if this population variance is defined as follows:

$$(12)\qquad \sigma^2_\lambda = \sum_i^N \lambda_i^2/(N - 1)\ .$$

Since we are interested in the quantity more usually defined as $\Sigma_i^N \lambda_i^2/N$, the estimate given by (11) must be corrected by a multiplicative factor of $(N - 1)/N$. Making this correction, and allowing the use of the same sigma, (11) becomes

$$(13)\qquad \hat{\sigma}^2_\lambda = \frac{n(N-1)}{NM(n-1)(m-1)}\,[m(M-1)s_y^2 - (M-m)\bar{s}_j^2]\ .$$

This formula will hold in general, whether or not the items are binary. When the items are binary, we note the following relationship:

$$(14) \qquad \bar{s}_j^2 = \bar{y}(1 - \bar{y}) - s_p^2 \ . \quad \text{(See Appendix 2, Proof 4.)}$$

Substituting (14) into (13) we obtain

$$(15) \qquad \hat{\sigma}_\lambda^2 = \frac{n(N-1)}{NM(n-1)(m-1)} \left[m(M-1)s_y^2 - (M-m)\{\bar{y}(1-\bar{y}) - s_p^2\}\right] ,$$

which is exactly the same as formula (2) given above by Lord for the case in which both examinee and item populations are finite.

For the case in which both N and M are infinite, we can either (a) take the limit of (15) as both N and M approach infinity or (b) note that the $E[MS]$ of Table 1 take the following forms and perform a derivation analogous to that from (9) to (15) above:

$$E[MS_E] = \sigma_\varepsilon^2 + m\sigma_\lambda^2$$

$$E[MS_I] = \sigma_\varepsilon^2 + n\sigma_\pi^2$$

$$E[MS_R] = \sigma_\varepsilon^2 \ .$$

In general, then, for infinite populations we have

$$(16) \qquad \hat{\sigma}_\lambda^2 = \frac{n}{(n-1)(m-1)} (ms_y^2 - \bar{s}_j^2) \ .$$

If the items are dichotomous, then

$$(17) \qquad \hat{\sigma}^2_\lambda = \frac{n}{(n-1)(m-1)} \left[ms^2_y - \bar{y}(1-\bar{y}) + s^2_p\right] .$$

This is exactly the same as Lord's formula (formula (3) above) for the infinite case.

Formulas for $\hat{\sigma}^2_\pi$ in the finite and infinite case are symmetric to those for σ^2_λ.

It can be easily argued that the foregoing model was unnecessarily restrictive in terms of the assumptions made. In the opening paragraph of his chapter on "item sampling" (Lord and Novick, 1968), Lord states that

> This chapter deals with the case where the ... test items are a random sample from the population of items. This item-sampling model makes no other assumptions about the nature of the test. (p. 234)

He later states that the examinees are a random sample from a population of examinees and makes the further assumption that the

> ... sample of items and the sample of examinees are drawn independently of each other. (p. 236)

In discussing the problems of estimation using this kind of model, Lord states that (author's symbols substituted for Lord's)

> In many of the usual, simple types of estimation problem, a population is

> completely specified by a convenient univariate frequency distribution. Many powerful estimation methods are available for such problems. Similarly a matrix population could be specified by an M-variate or an N-variate frequency distribution, provided that an appropriate and convenient mathematical form could be found... .
>
> In the absence of an adequate parametric form for a frequency distribution, how are we to describe a matrix population without using a huge number of parameters? (pp. 237-238)

Lord goes on to answer this question with the presentation of Hooke's formulation of matrix sampling, which rests in fact only on the assumption of independent and random sampling of examinees and items.

It is well known that point estimates of the variance components for the model given by (4) do not require any assumption regarding the shape of the distributions (see footnote, p. 17). In fact, this parametric model <u>with several modifications</u> can be used with the same, weak assumptions of the Hooke approach. This can best be seen by gradually relaxing the restrictions put on (4) for deriving the expected mean squares.

One severe restriction in the model was the assumption of additivity, viz., the lack of provision for an interaction effect apart from error. Eliminating this assumption, the model given by (4) can be slightly modified to produce the following linear and <u>nonadditive</u> model:

$$X_{ij} = \mu + \lambda_i + \pi_j + \lambda\pi_{ij} + \varepsilon_{ij} \qquad (18)$$

where

$\lambda\pi_{ij}$ = an additional independently distributed random variable representing the interaction effect between examinee i and item j in the matrix population.

We now possess a model where the form of the error-free item covariance matrix among items in the population is arbitrary, i.e., there is no necessity for assuming the matrix to be of rank 1.

But with the addition of an interaction effect, the unrealistic assumption of its independence of examinee and item effects is made. Furthermore, the restriction of homogeneity of error variance is still present. Cornfield and Tukey (1956) have demonstrated that expected values of mean squares can be derived using what is referred to as a "pigeonhole" model. Specifically, for the two-way classification the authors describe the model as follows (corresponding symbols of the present paper are substituted for those of Cornfield and Tukey):

> Let there be NM pigeonholes arranged in N rows and M columns. Let there be at least R elements in the population in each pigeonhole. Let a sample of n rows be drawn from the N potential rows. Let a sample of m columns be drawn from the M potential columns. The nm intersections of a selected row with a selected column specify the nm pigeonholes which become the cells of the actual experiment. In each of these nm cells, let a sample of r elements be drawn. The values of the nmr elements thus obtained are the numbers which are to be analyzed. Assume that all the

> samplings -- of rows, of columns, and within pigeonholes -- are at random and independent of one another. This is the only assumption we shall make. Note that it is an assumption about the set-up of the experiment and not about the behavior of those things on which the experiment is performed. (p. 909)

They go on to point out the generality of their model in that (a) no constant variance is assumed for the cells and (b) no assumption is made about interaction, i.e., the interaction effects are dependent upon the particular rows and columns that happen to be sampled.

Now consider Table 2 which presents the most general form of the E[MS] for this model. When $r = 1$ in this pigeonhole model (see Table 3), we have the matrix sampling situation, but based only on the above assumptions.

At this point, a subtle and important conceptual problem arises. How are we to treat the MN "populations" of replications of size R, given that r will always equal one? Three distinct choices are available: (a) $R = \infty$, (b) R finite and greater than r, or (c) R finite and $R = r = 1$. We must also be concerned with our treatment of M (as finite or infinite) since it enters into the $E[MS_E]$. From Table 4 it is evident that (a) when M is finite, R must be finite and equal to one in order that $\hat{\sigma}^2_\lambda$ be computed exactly and (b) when M is infinite, the exact estimate of σ^2_λ can always be computed regardless of the value of R. (In Table 4 both M and N are treated alike to illustrate the analogy for estimating σ^2_π.)

In any case, it is clear that the explicit use of analysis of variance estimation procedures can be used to derive formulas

TABLE 2

General Form of the E[MS] for the Two-Factor Design

Source	E[MS]
E	$(1 - r/R)\sigma_\varepsilon^2 + r(1 - m/M)\sigma_{\lambda\pi}^2 + rm\sigma_\lambda^2$
I	$(1 - r/R)\sigma_\varepsilon^2 + r(1 - n/N)\sigma_{\lambda\pi}^2 + rn\sigma_\pi^2$
EI	$(1 - r/R)\sigma_\varepsilon^2 + r\sigma_{\lambda\pi}^2$
Error	$(1 - r/R)\sigma_\varepsilon^2$

TABLE 3

The E[MS] of Table 2 when r = 1

Source	E[MS]
E	$(1 - 1/R)\sigma_\varepsilon^2 + (1 - m/M)\sigma_{\lambda\pi}^2 + m\sigma_\lambda^2$
I	$(1 - 1/R)\sigma_\varepsilon^2 + (1 - n/N)\sigma_{\lambda\pi}^2 + n\sigma_\pi^2$
EI	$(1 - 1/R)\sigma_\varepsilon^2 + \sigma_{\lambda\pi}^2$

TABLE 4

E[MS] of Table 3 for Indicated Values of M and R

M (and N) finite

Source	E[MS]
R infinite	
E	$\sigma_\varepsilon^2 + (1 - m/M)\sigma_{\lambda\pi}^2 + m\sigma_\lambda^2$
I	$\sigma_\varepsilon^2 + (1 - n/N)\sigma_{\lambda\pi}^2 + n\sigma_\pi^2$
EI	$\sigma_\varepsilon^2 + \sigma_{\lambda\pi}^2$
R finite; R > r	
E	$(1 - 1/R)\sigma_\varepsilon^2 + (1 - m/M)\sigma_{\lambda\pi}^2 + m\sigma_\lambda^2$
I	$(1 - 1/R)\sigma_\varepsilon^2 + (1 - n/N)\sigma_{\lambda\pi}^2 + n\sigma_\pi^2$
EI	$(1 - 1/R)\sigma_\varepsilon^2 + \sigma_{\lambda\pi}^2$
R finite; R = 1	
E	$(1 - m/M)\sigma_{\lambda\pi}^2 + m\sigma_\lambda^2$
I	$(1 - n/N)\sigma_{\lambda\pi}^2 + n\sigma_\pi^2$
EI	$\sigma_{\lambda\pi}^2$

TABLE 4 (continued)

M (and N) infinite

Source	E[MS]
E	$\sigma_\varepsilon^2 + \sigma_{\lambda\pi}^2 + m\sigma_\lambda^2$
I	$\sigma_\varepsilon^2 + \sigma_{\lambda\pi}^2 + n\sigma_\pi^2$
EI	$\sigma_\varepsilon^2 + \sigma_{\lambda\pi}^2$

R finite; R > r

E	$(1 - 1/R)\sigma_\varepsilon^2 + \sigma_{\lambda\pi}^2 + m\sigma_\lambda^2$
I	$(1 - 1/R)\sigma_\varepsilon^2 + \sigma_{\lambda\pi}^2 + n\sigma_\pi^2$
EI	$(1 - 1/R)\sigma_\varepsilon^2 + \sigma_{\lambda\pi}^2$

R finite; R = 1

E	$\sigma_{\lambda\pi}^2 + m\sigma_\lambda^2$
I	$\sigma_{\lambda\pi}^2 + n\sigma_\pi^2$
EI	$\sigma_{\lambda\pi}^2$

equivalent to those of Lord without making any stronger assumptions. Although the derivations for the variance estimate in this section were based on the strong model given by (4), exactly the same algebra would be involved for any of the models in Table 4. However, the usual procedure of taking the expected value of both sides of (18) to obtain an estimate of μ cannot be done in view of the weak assumptions made (e.g., $\lambda\pi_{ij}$ was not assumed to be statistically independent of the remaining effects). It can easily be shown, however, that $\hat{\mu} = \bar{X}$ under the sampling assumptions made (see Appendix 2, Proof 6).

4. DISCUSSION

This section will attempt to coordinate the theoretical development of matrix sampling given in the previous chapter with the actual use of, and problems with, the technique when applied to psychometric data. Specifically, the discussion will center around the use of multiple matrix sampling, the possibility of obtaining negative variance estimates being given particular attention.

Multiple matrix sampling (briefly discussed in the first section) is the process of randomly drawing more than one matrix sample from a matrix population, computing the desired parameter estimate from each sample, and combining these estimates to produce one final, more stable estimate. There are at least three ways in which sampling from the item population can be systematically accomplished: (a) sampling with replacement, i.e., any given item sample can be drawn more than once, (b) sampling with "restricted replacement," i.e., any particular item sample cannot be drawn more than once but any given item can appear in more than one item sample, or (c) sampling without replacement, i.e., no item or item sample can be drawn more than once. Since the same remarks apply to the sampling of examinees, there is a total of nine different ways to draw matrix samples from the matrix population.

Unfortunately, Lord's discussion of multiple matrix sampling is rather vague from both theoretical and methodological standpoints.

In fact, the only discussion of multiple matrix sampling approaching some degree of rigor deals only with estimating the population mean (Lord and Novick, 1968). The author can find no such discussion regarding the estimation of the population variance; yet the procedures of multiple matrix sampling have been employed for both estimates, starting with Lord (1962). To be more specific, in section 11.12 (Lord and Novick, 1968) Lord offers (with no formal proof) the following statements (corresponding symbols of the present paper have been used in place of Lord's):

> The methods of the preceding section [estimating a mean from a single matrix sample] do not make full use of the advantages of item sampling. Under a more efficient procedure to be outlined in this section, the examiner administers different samples of binary items to different subgroups of examinees. This procedure draws on a mathematical formulation that has arisen from certain unpublished suggestions of Dr. William W. Turnbull.
>
> Suppose K nonoverlapping random samples of m binary items each are drawn (without replacement) from an M-item test and treated as separate subtests; it is not required that $K = M/m$ or $K = N/n$. A different subtest is administered to each of K nonoverlapping random samples of n examinees drawn from a population of N examinees. If $\bar{X}_k$ is the mean relative score of subgroup k on an m-item test, then the average $\bar{X}_k$ is an unbiased estimater of μ, the mean score of the N examinees on the M-item test. (p. 255)

No parallel theorem is stated for variance estimates, yet the above procedure has been used for the $\hat{\sigma}^2_\lambda$ as well as the $\bar{X}_k$ of the matrix samples (see Chapter 2 starting with Lord, 1962). Consider,

however, the following statements in section 11.14 (Lord and Novick, 1968) regarding the estimation of σ_λ^2 for finite populations:

> As we saw in section 11.12, it is much better to administer many different m-item subtests than just one. If this is done, it is again important ... that every item appear an equal number of times; that all [possible] pairs of items appear in the subtests, if possible; and that each pair be administered to the same number of examinees. When all [possible] pairs can not be used, good balanced designs may sometimes be found with the help of tables of balanced incomplete blocks.... (p. 259)

(Knapp (in press) gives a detailed discussion of the application of balanced incomplete block designs to the estimation of the mean and variance.) Clearly, if these above criteria are satisfied, then it is impossible to have <u>nonoverlapping</u> samples of items.

The above conflicting points of view and practice must further be reconciled with the following statements (Lord, 1962 -- see discussion of this paper in Chapter 2):

> In retrospect, the foregoing item-sampling procedure is seen to have been unnecessarily inefficient. Items were sampled with replacement after each sampling for the reason that such sampling is effectively the same as sampling from an infinite pool of items, and the available formulas in Lord (1960) for utilizing the resulting data are discussed in terms of sampling from an infinite pool. It would have been better to sample without replacement, thus dividing the 70 items at random into 10 overlapping 7-item tests. Hooke's (1956a) basic derivations show that the same formulas would be valid for such sampling without replacement. (pp. 261-262)

The author can find no discussion of multiple matrix sampling in Hooke (1956a or b). Perhaps Lord was simply referring to the fact that theory was available for sampling a matrix sample when the examinee and/or item population is finite. As seen in section 3, the Cornfield and Tukey (1956) approach supplies the same finite sampling theory. In both presentations, the process of selecting more than one matrix sample is discussed only in the context of defining the "inherited on the average" property. For example, consider the following definition by Hooke (1956a):

> Let x_I $(I = 1,2,\ldots,N)$ be any population of N numbers, and let x_i $(i = 1, 2,\ldots,n)$ represent elements of a sample of size n from this population. Let $f(n; x_1,\ldots,x_n)$ be a polynomial which is symmetric in the x_i and has coefficients which are functions of n. Such a function extends obviously to a polynomial $f(N; x_1,\ldots,x_N)$, the coresponding symmetric polynomial in the x_I, with the coefficients changed only by replacing n by N. Writing "ave" for the operation of averaging over all $\binom{N}{n}$ distinct samples of size n from the population, we say that $f(n; x_1,\ldots,x_n)$ is 'inherited on the average' if
>
> $$\text{ave } f(n; x_1,\ldots,x_n) = f(N; x_1,\ldots,x_N) .$$
>
> (p. 55)

Clearly, this type of sampling is the second type referred to above as sampling with restricted replacement. It would seem appropriate to stick to this type of sampling if estimates from multiple samples

were to remain unbiased. In this case, nonoverlapping sampling (the sampling of matrices without replacement) would be an invalid restriction.

This point can perhaps be better illustrated from the analysis of variance view point and the Cornfield and Tukey (1956) approach. In order to derive expected values of mean squares, the only assumptions made were that rows and columns were randomly and independently sampled from their corresponding populations. In multiple nonoverlapping matrix samples, only the first sample satisfies these assumptions. The fact that these rows and columns (examinees and items) cannot be included in subsequent matrix samples imposes a dependence and non-randomness on the rows and columns sampled in subsequent matrix samples.

The author does not know what effect the restriction of non-overlapping matrix samples has on the resulting parameter estimates. Although the kth $(k > 1)$ matrix sample will not strictly conform to the above sampling assumptions, it might be argued intuitively that the items and examinees of this sample are unbiased in the sense that they would have had the same chance of being selected at the outset as those of any other sample. The parameter estimates might then be considered in the same sense as being unbiased.

The reason for heavy concentration on the theory thus far lies in the following fact: It is possible that the variance estimate generated from the matrix sampling formula is negative. Recognition of the fact that variance component estimates can be negative is not

unique to the present paper. Thompson (1962) in regard to variance component estimates makes the following statements:

> The traditional estimators ... are obtained ... by equating the mean-squares to their expectations and solving. Clearly the traditional estimate ... may be negative; should this occur, we do not believe that any such statistical analysis would become useful until it is decided what to do with the negative estimate. This, then, is an example of what we mean by 'the problem of negative estimates of variance components'. Two possible explanations of a negative estimate present themselves: (1) the assumed model may be incorrect and (2) statistical noise may have obscured the underlying physical situation. (p. 274)

Husek and Sirotnik (1968) actually obtained a negative variance estimate while conducting a study (see section 2) involving multiple matrix sampling from an already known matrix population of data. By setting Lord's formula less than zero and simplifying the resulting inequality, they showed that a negative variance estimate would be obtained whenever

$$ms_y^2 < s_x^2 - s_p^2$$

or, alternatively, whenever

$$\alpha < 0$$

where α is Cronbach's (1951) generalized coefficient of internal consistency. (If the items are dichotomous, α = K - R 20, the Kuder-Richardson (1937) coefficient.)

In the present framework, this result is relatively trivial. Hoyt (1943) showed that the repeated measures analysis of variance design could be used as an alternative approach to obtaining a measure of internal consistency equivalent to that of K - R 20. (See Appendix 2, Proof 5 for a derivation of this fact for the more general case of Cronbach's coefficient alpha.) In the most general form, Hoyt's result can be written as follows:

$$\alpha = (MS_E - MS_{IE})/MS_E \quad . \tag{19}$$

By substituting the E[MS] for the MS in this equation, it can easily be seen that the ratio is a ratio of true score variance to the total true score plus error variance. Although the above formula was originally derived using the strong analysis of variance model first presented in section 3, it is just as valid using the weakest model, viz., the Cornfield and Tukey (1956) pigeon-hole approach. This point was made by Cronbach, Rajaratnam, and Gleser (1963) who attempted to free reliability theory from the concept of "parallel" measures. They redefined reliability in terms of generalizing from a sample of observations to a universe (or population) of observations. They did not wish to be restrained by any statistical characteristics of the item (or examinee) population (e.g., unit rank, no systematic examinee - by - item interaction, equality of error variance); hence, the Cornfield and Tukey approach provided the needed theoretical framework.

Now, returning to (19), it is clear that whenever $MS_E < MS_{IE}$, $\alpha < 0$. If we substitute the equivalence relations given in section 3

for the sums of squares into this inequality, we obtain

$$(20) \qquad \frac{nms_y^2}{n-1} < \frac{nms_j^{-2} - nms_y^2}{(n-1)(m-1)}$$

$$(m-1)s_y^2 < \bar{s}_j^2 - s_y^2$$

$$ms_y^2 < s_x^2 - s_p^2 ,$$

which is exactly the relationship found by Husek and Sirotnik (1968). The relationship between $\hat{\sigma}_\lambda^2$ and α can be seen directly by first noting that

$$(21) \qquad \hat{\sigma}_\lambda^2 = \frac{MS_E - MS_{IE}}{m}$$

and then combining (19) and (21) yielding

$$(22) \qquad \hat{\sigma}_\lambda^2 = \alpha \frac{MS_E}{m} .$$

The point, again, is this: Since σ_λ^2 (<u>not</u> $\hat{\sigma}_\lambda^2$) can never be negative; when $\hat{\sigma}_\lambda^2 < 0$, then either (a) the theoretical assumptions underlying the derivation of the E[MS] have been violated or (b) we have been victimized by extreme sampling fluctuation. With respect to the first explanation, the literature is not clear on what the effect (if any) is on variance component estimates when matrices are sampled without replacement. Lord states (Lord and Novick, 1968) that the estimates of the population mean from such

samples are unbiased. He implies that the same holds true for estimates of the population variance (Lord, 1962).

Suppose the second explanation is accepted. That is, suppose a researcher is using multiple matrix sampling to establish norms on some test and he obtains a negative variance estimate from one or more matrix samples. The researcher is in the position of having to deal with these negative values in some kind of averaging process to arrive at a final estimate. Consider the following three possible approaches, presented in decreasing order in terms of mathematical justification and increasing order (in the opinion of the author) in terms of reasonability.

It is common practice to regard negative variance component estimates as, for all practical purposes, zero. Mathematical justification for this practice stems from the fact that the maximum likelihood estimate of σ^2_λ is zero when $MS_{IE} > MS_E$ under the restriction of positive component estimates (Thompson, 1962). But consider the following remarks by Scheffé (1959):

> It may happen with positive probability that the estimate of a variance component is negative ... Since the estimated parameter is nonnegative, the estimate is sometimes modified by redefining it to be zero when it is negative ... We prefer <u>not</u> to use such modified estimates: their distribution theory is more complicated ... and the modified estimates are biased. (p. 229)

Sometimes when the researcher is willing to break away from his search for the mathematically rigorous solution, he can stumble upon

more non-rigorous data analytic methods which might correspond more closely to the behavior of the real world. (Such correspondence would, of course, have to be validated empirically.) As Tukey (1962) puts it in his discussion of the future of data analysis,

> We should seek out unfamiliar summaries of observational material, and establish their useful properties.... Many seem to find it essential to begin with a probability model containing a parameter, and then to ask for a good estimate for this parameter (too often, unfortunately, for one that is optimum). Many have forgotten that data analysis can, sometimes quite appropriately, precede probability models, that progress can come from asking what a specified indicator (= a specified function of the data) may reasonably be regarded as estimating. Escape from this constraint can do much to promote novelty. (p. 5)

Lindquist (1956) states that variance component estimates are approximately normally distributed when the degrees of freedom involved are very large. Clearly, in matrix sampling n and nm are apt to be fairly small. The author knows of no research regarding the small sampling distribution of variance component estimates. The negative variance estimates in matrix sampling data, however, suggest that at least one end of this distribution is rather long-tailed. Making the assumption that this distribution is approximately symmetrical, the other end can be regarded as long-tailed, produced by extremely high component estimates. In other words, in a distribution of variance component estimates obtained from small samples, any negative estimates (and an equal number

of largest positive estimates) might be regarded as "outliers" and not representative of the population parameter.

A simple and intuitive procedure for handling these outliers is to "trim" or "Winsorize" (Tukey, 1962) the distribution of estimates before averaging to produce the final estimate. A <u>trimmed</u> distribution is one where an equal number of lowest and highest outliers are eliminated from the distribution. In a <u>Winsorized</u> distribution, these outliers are forced equal to the remaining lowest and highest observations respectively.

Specifically, suppose k ordered variance estimates e_i have been obtained by multiple matrix sampling and t of these are negative. Then the mean of the Winsorized distribution of these estimates would be given as follows:

$$\bar{e} = \frac{1}{k}\left(te_{(t+1)} + \sum_{i=t+1}^{k-t} e_i + te_{(k-t)} \right).$$

Just what the shape of the sampling distribution of variance component estimates is must be settled empirically, starting with computer simulated data and corresponding distributions of mean variance component estimates using both maximum likelihood and Winsorizing approaches. Depending upon the shape of obtained distributions of variance component estimates for this type of small sampling, various non-symmetrical Winsorizing approaches (Dixon, 1960) might also be compared.

The third approach to be suggested will be prefaced by the

following discussion regarding confidence intervals for variance component estimates by Scheffé (1959):

> Some discussion is required because one or both end points of the interval may be negative while the true value of [the parameter] is of course nonnegative...
>
> It would be mathematically correct to modify the interval so that if the left end point is negative it is replaced by zero and if the right end point is negative it is also replaced by zero...
>
> Although there is nothing in the formal theory of confidence intervals to justify it, most users of confidence intervals have a more or less conscious feeling that the length of a two-sided confidence interval is a measure of the error of some point estimate of the parameter...
>
> In light of the above discussion we see that if the interval is considerably shortened by deleting the part, if any, to the left of the origin, a misleading impression of the accuracy of the estimation may result. If the interval is completely to the left of the origin one might consider translating it until it just includes the origin, to meet the above objection to shortening it. However, one might again feel on nonmathematical and intuitive grounds that an interval estimate like that from -5 to -3 is stronger evidence that the true value of a nonnegative parameter is zero than that from -2 to 0. (pp. 229-231)

Extrapolating from Scheffé's argument to the situation in multiple matrix sampling, one might intuitively feel that the magnitude of variance component estimates carries with it important information -- regardless of whether it is positive or negative. Thus, it would seem reasonable to average these estimates without any modifications whatsoever.

Table 5 presents the ten matrix sample estimates obtained in the Husek and Sirotnik (1968) study (see section 2). Averages obtained under the three possible ways of handling the negative estimate are also presented. It can be seen that equating the negative estimate to zero or symmetrical Winsorizing produce nearly the same result. Averaging in the negative estimate, however, yields a final estimate which is substantially closer to the actual population value. Again, a simulation study is needed to evaluate the relative merits of the three proposed procedures.

TABLE 5

A Comparison of Three Alternative Procedures for Handling Negative Variance Component Estimates*
Data taken from Husek and Sirotnik (1968)

Matrix Sample	Obtained Estimates	Equating Negative Estimate To Zero	Symmetrical Winsorizing
1	.00645	.00645	.00583
2	.00583	.00583	.00583
3	.00479	.00479	.00479
4	.00412	.00412	.00412
5	.00384	.00384	.00384
6	.00276	.00276	.00276
7	.00191	.00191	.00191
8	.00179	.00179	.00179
9	.00074	.00074	.00074
10	-.00181	0	.00074
Average	.00304	.00322	.00323

(Population variance = .00258)

*Variance estimates are of examinee mean scores.

REFERENCES

Cahen, L. S., Romberg, T. A., & Zwirner, W. The estimate of mean achievement scores for schools by the item-sampling technique. Paper presented at the meeting of the Psychometric Society, 1967.

Cook, D. L., & Stufflebeam, D. L. Estimating test norms from variable size item and examinee samples. Journal of Educational Measurement, 1967, 4, 27-33. (a)

Cook. D. L., & Stufflebeam, D. L. Estimating test norms from variable size item and examinee samples. Educational and Psychological Measurement, 1967, 27, 601-610. (b)

Cornfield, J., & Tukey, J. W. Average values of mean squares in factorials. Annals of Mathematical Statistics, 1956, 4, 907-949.

Cronbach, L. J. Coefficient alpha and the internal structure of tests. Psychometrika, 1951, 16, 297-334.

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. Theory of generalizability: A liberation of reliability theory. British Journal of Statistical Psychology, 1963, 16, 137-163.

Dixon, W. J. Simplified estimation from censored normal samples. Annals of Mathematical Statistics, 1960, 31, 385-391.

Gulliksen, H. Theory of mental tests. New York: Wiley, 1950.

Hooke, R. Symmetric functions of a two-way array. Annals of Mathematical Statistics, 1956, 27, 55-79. (a)

Hooke, R. Some applications of bipolykays to the estimation of variance components and their moments. Annals of Mathematical Statistics, 1956, 27, 80-98. (b)

Horst, P. Psychological measurement and prediction. Belmont, Calif.: Wadsworth, 1966.

Hoyt, C. Test reliability estimated by analysis of variance. Psychometrika, 1941, 6, 153-160.

Husek, T. R., & Sirotnik, K. Matrix sampling in educational research: An empirical investigation. Paper presented at the 1968 convention of the American Educational Research Association.

Knapp, T. R. An application of balanced incomplete block designs to the estimation of test norms. Educational and Psychological Measurement. in press.

Lindquist, E. F. Design and analysis of experiments in psychology and education. Boston: Houghton Mifflin, 1956.

Lord, F. M. Sampling fluctuations resulting from the sampling of test items. Psychometrika, 1955, 20, 1-22.

Lord, F. M. Statistical inferences about true scores. Psychometrika, 1959, 24, 1-17. (a)

Lord, F. M. An approach to mental test theory. Psychometrika, 1959, 24, 283-302. (b)

Lord, F. M. Use of true-score theory to predict moments univariate and bivariate observed-score distributions. Psychometrika, 1960, 25, 325-342.

Lord, F. M. Estimating norms by item sampling. Educational and Psychological Measurement, 1962, 22, 259-267.

Lord, F. M. Item sampling in test theory and research design. Princeton, N. J.: Educational Testing Service, Rb-65-22, 1965.

Lord, F. M., & Novick, M. R. Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968.

Magnusson, D. Test theory. Reading: Addison-Wesley, 1967.

Plumlee, L. B. Estimating means and standard deviations from partial data--an empirical check on Lord's item sampling technique. Educational and Psychological Measurement, 1964, 24, 623-630.

Scheffé, H. The analysis of variance. New York: John Wiley & Sons, 1959.

Thompson, W. A. The problem of negative estimates of variance components. Annals of Mathematical Statistics, 1962, 33, 273-289.

Tukey, J. W. The future of data analysis. Annals of Mathematical Statistics, 1962, 33, 1-67.

Winer, B. J. Statistical principles in experimental design. New York: McGraw-Hill, 1962.

APPENDIX 1*

Notation

α . . . Cronbach's generalized coefficient of internal consistency among items.

ε_{ij} . . . Error effect in the examinee-by-item analysis of variance design (see p. 15).

E[ ] . . . Expected value operator.

k . . . Subscript denoting the k^{th} matrix sample in multiple matrix sampling.

K . . . Number of multiple matrix samples.

K-R 20 . . . Kuder-Richardson coefficient of internal consistency among dichotomous items.

λ_i . . . Examinee effect in the examinee-by-item analysis of variance design (see p. 15).

$\lambda\pi_{ij}$ . . . Interaction effect in the examinee-by-item analysis of variance design (see p. 23).

μ, $\hat{\mu}$ . . . Mean and estimated mean of the matrix population.

m, M . . . Sample and population sizes of the items.

n, N . . . Sample and population sizes of the examinees.

π_j . . . Item effect in the examinee-by-item analysis of variance design (see p. 15).

p_j, $\bar{p}$ . . . Sample item j mean $(\bar{X}._j)$ and mean of the p_j.

r, R . . . Sample and population sizes of the cells in a 2-factor analysis of variance design.

σ^2_λ, $\hat{\sigma}^2_\lambda$ . . . Variance and estimated variance of the λ_i.

σ^2_π, $\hat{\sigma}^2_\pi$ . . . Variance and estimated variance of the π_j.

* This appendix is to be used in conjunction with Figure 2.

$\sigma^2_{\lambda\pi}$ . . . Variance of the $\lambda\pi_{ij}$.

σ^2_{ε} . . . Variance of the ε_{ij}.

s^2_y . . . Variance of the y_i.

s^2_p . . . Variance of the p_j.

s^2_j, $\bar{s}^2_j$ . . . Variance of item j (computed over examinees) and average of the s^2_j.

X_{ij}, $\bar{X}$ . . . The response of examinee i to item j and the mean of the X_{ij}.

y_i, $\bar{y}$ . . . Sample examinee mean ($\bar{X}_i$.) and the mean of the y_i.

APPENDIX 2

Proofs

Proof 1

$$SS_E = m \sum_i (y_i - \bar{X})^2 \quad \text{as defined in Table 1}$$

$$= mn \sum_i (y_i - \bar{X})^2/n$$

$$= mns_y^2 \quad \text{by definition of a variance}$$

Proof 2

$$SS_I = n \sum_j (p_j - \bar{X})^2 \quad \text{as defined in Table 1}$$

$$= nm \sum_j (p_j - \bar{X})^2/m$$

$$= nms_p^2 \quad \text{by definition of a variance}$$

Proof 3

$$SS_R = \sum_i \sum_j (X_{ij} - y_i - p_j + \bar{X})^2 \quad \text{as defined in Table 1}$$

$$= \sum_i \sum_j [(X_{ij} - p_j) - (y_i - \bar{X})]^2$$

$$= \sum_i \sum_j (X_{ij} - p_j)^2 + \sum_i \sum_j (y_i - \bar{X})^2$$

$$- 2 \sum_i \sum_j (X_{ij} - p_j)(y_i - \bar{X})$$

But

$$\sum_i \sum_j (X_{ij} - p_j)^2 = \sum_j n \sum_i (X_{ij} - p_j)^2/n$$

$$= n \sum_j s_j^2$$

$$= nms_j^2 ,$$

$$\sum_i \sum_j (y_i - \bar{X})^2 = \sum_j n \sum_i (y_i - \bar{X})^2$$

$$= \sum_j ns_y^2$$

$$= nms_y^2 ,$$

and

$$-2 \sum_i \sum_j (X_{ij} - p_j)(y_i - \bar{X}) = -2 \sum_i (y_i - \bar{X}) \sum_j (X_{ij} - p_j)$$

$$= -2 \sum_i (y_i - \bar{X}) \Big(\sum_j X_{ij} - \sum_j p_j \Big)$$

$$= -2 \sum_i (y_i - \bar{X})(my_i - m\bar{X})$$

$$= -2m \sum_i (y_i - \bar{X})^2$$

$$= -2nms_y^2 .$$

Substituting,

$$SS_R = nm\bar{s}_j^2 + nms_y^2 - 2nms_y^2$$

$$= nm\bar{s}_j^2 - nms_y^2$$

Proof 4

$$\bar{s}_j = \sum p_j(1 - p_j)/n \quad \text{for binary items}$$

$$= \sum p_j/n - \sum p_j^2/n$$

$$= \bar{X} - \sum p_j^2/n$$

But

$$s_p^2 = \sum p_j^2/n - \bar{X} .$$

or, rearranging terms,

$$\sum p_j^2/n = s_p^2 + \bar{X}^2 .$$

Substituting,

$$\bar{s}_j^2 = \bar{X} - \bar{X}^2 - s_p^2$$

$$= \bar{X}(1 - \bar{X}) - s_p^2 \quad \text{or} \quad \bar{y}(1 - \bar{y}) - s_p^2$$

Proof 5

$$\frac{MS_E - MS_R}{MS_E} = \frac{\dfrac{nms_y^2}{n-1} - \dfrac{n\bar{m}s_j^2 - nms_y^2}{(n-1)(m-1)}}{\dfrac{nms_y^2}{n-1}}$$

$$= \frac{(m-1)s_y^2 - \bar{s}_j^2 + s_y^2}{(m-1)s_y^2}$$

$$= \frac{ms_y^2 - \bar{s}_j^2}{(m-1)s_y^2}$$

$$= \frac{ms^2_{my} - m^2\bar{s}^2_j}{(m-1)s^2_{my}} \qquad \text{(where } s^2_{my} = \text{variance of examinee } \underline{\text{total}} \text{ scores)}$$

$$= \frac{m}{m-1}\left(1 - \frac{m\bar{s}^2_j}{s^2_{my}}\right)$$

$$= \frac{m}{m-1}\left(1 - \frac{\Sigma s^2_j}{s^2_{my}}\right)$$

$$= \text{Cronbach's coefficient alpha}$$

Proof 6

To show that $\hat{\mu} = \bar{X}$, it must be shown that the mean $(\bar{\bar{X}})$ of the sampling distribution of $\bar{X}$ (generated under the sampling assumptions of the model) is equal to μ.

The sampling distribution is made up of the means of $\binom{N}{n}\binom{M}{m}$ possible $n \times m$ matrix samples from the $N \times M$ matrix population.

Each X_{ij} will appear in $\binom{N-1}{n-1}\binom{M-1}{m-1}$ matrix samples.

For any given matrix sample,

$$\bar{X} = \frac{1}{nm}(\text{sum } X_{ij} \text{ in that sample}) .$$

Over all possible matrix samples, then,

$$\bar{\bar{X}} = \frac{1}{\binom{N}{n}\binom{M}{m}} \sum \bar{X}$$

$$= \frac{1}{\binom{N}{n}\binom{M}{m}} \cdot \frac{1}{nm} \text{ (sum } X_{ij} \text{ in } \underline{\text{all}} \text{ samples)}$$

$$= \frac{1}{\binom{N}{n}\binom{M}{m}} \cdot \frac{1}{nm} \binom{N-1}{n-1}\binom{M-1}{m-1} \sum \sum X_{ij}$$

$$= \frac{1}{NM} \sum \sum X_{ij}$$

$$= \mu$$

APPENDIX 3

A Brief Introduction to Lord's Use of Hooke's Approach to Derive Formulas for the Mean and Variance Estimates in Matrix Sampling

Let $\{i_1, i_2, i_3, \ldots, i_i, \ldots, i_a\}$ be a set of a alternative indices for examinees.

Let $\{j_1, j_2, j_3, \ldots, j_j, \ldots, j_b\}$ be a set of b alternative indices for items.

Let $\{p_1, p_2, p_3, \ldots\ldots\ldots\ldots, p_{ab}\}$ be a set of ab integral powers.

Let X_{ij} be the response of examinee i to item j in the n-examinee by m-item matrix randomly sampled from a population N-examinee by M-item matrix. (For the following discussion, both N and M will be taken to be infinite.)

Definition: Denoting a generalized symmetric mean as gsm,

$$gsm = \frac{1}{T} \sum_{\neq}^{T} (X_{i_1 j_1}^{p_1} X_{i_1 j_2}^{p_2} \cdots X_{i_1 j_b}^{p_b})(X_{i_2 j_1}^{p_{b+1}} X_{i_2 j_2}^{p_{b+2}} \cdots X_{i_2 j_b}^{p_{2b}}) \cdots$$

$$(X_{i_a j_1}^{p_{b(a-1)+1}} X_{i_a j_2}^{p_{b(a-1)+2}} \cdots X_{i_a j_b}^{p_{ab}})$$

where

$$T = n(n-1) \cdots (n-a+1)m(m-1) \cdots (m-b+1)$$

and

$\neq$ denotes the distinctness of the i_i and the j_j .

Note: The gsm is symmetric, i.e., its value is invariant under permutations of rows and/or columns of the matrix in X_{ij}.

Definition: A bipolykay is a linear combination of gsm's.

Theorem: A gsm is inherited on the average, i.e.,

E[gsm in matrix sample] = E[gsm in matrix population] .

Corollary: A bipolykay is also inherited on the average.

For convenience, we can specify any given gsm by an "operator matrix" whose rows specify the a alternative examinee indices, columns specify the b alternative item indices, and elements specify the ab integral powers for the corresponding $X_{i_i j_j}$. Thus, the gsm as defined above can be specified as follows:

$$\begin{bmatrix} p_1 & p_2 & \cdots & p_j & \cdots & p_b \\ p_{b+1} & p_{b+2} & \cdots & p_{b+j} & \cdots & p_{2b} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ p_{b(i-1)+1} & p_{b(i-1)+2} & \cdots & p_{b(i-1)+j} & \cdots & p_{ib} \\ p_{b(a-1)+1} & p_{b(a-1)+2} & \cdots & p_{b(a-1)+j} & \cdots & p_{ab} \end{bmatrix}$$

There is only one 1^{st} power gsm, namely (zero rows and columns are used to achieve uniformity in notation for 1^{st} and 2^{nd} degree gsm's),

$$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} = \frac{1}{nm} \sum_{i}^{n} \sum_{j}^{m} X_{ij} = \bar{X}$$

There are four possible 2^{nd} degree gsm's, for example

$$\begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix} = \frac{1}{nm} \sum_{i}^{n} \sum_{j}^{m} X_{ij}^2$$

$$\begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix} = \frac{1}{m(m-1)n} \sum_{i}^{n} \sum_{j_1 \neq j_2}^{n(m-1)} X_{ij_1} X_{ij_2}$$

$$= \frac{1}{m-1} (ms_y^2 + n\bar{y}^2 - \bar{y})$$

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \frac{1}{m(m-1)n(n-1)} \sum_{i_1 \neq i_2}^{n(n-1)} \sum_{j_1 \neq j_2}^{m(m-1)} X_{i_1 j_1} X_{i_2 j_2}$$

$$= \frac{1}{(n-1)(m-1)} [(nm - n - m)\bar{y}^2 - ns_p^2 - ms_y^2 + \bar{y}]$$

Applying the expected value theorem for gsm's, it can be shown that

$$E\begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix} = \sigma_\lambda^2 + \mu^2$$

and

$$E\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \mu^2 \ .$$

Therefore (by the above corollary), the bipolykay equal to the difference of the above two gsm's has the following expected value:

$$E\left[\begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix} - \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right] = \sigma_\lambda^2 \ .$$

Hence,

$$\hat{\sigma}_\lambda^2 = \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix} - \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$= \frac{n}{(n-1)(m-1)} \left[ns_y^2 - \bar{y}(1 - \bar{y}) + s_p^2\right] \ .$$

Referring back to the 1^{st} degree gsm, it is clear that

$$\hat{\mu} = \bar{\bar{X}} \ .$$